Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Computer Science 17 (2013) 803 – 811

Procedia Computer Science

Information Technology and Quantitative Management (ITQM2013)

Advertisement Click-Through Rate Prediction using Multiple Criteria Linear Programming Regression Model

Fang Wang[a]*, Warawut Suphamitmongkol[a], Bo Wang[a]

[a]*Research Center on Fictitious Economy & Data Science, Chinese Academy of Sciences, Beijing 100190, China*

Abstract

In advertisement industry, it is important to predict potentially profitable users who will click target ads (i.e., Behavioral Targeting). The task selects the potential users that are likely to click the ads by analyzing user's clicking/web browsing information and displaying the most relevant ads to them. In this paper, we present a Multiple Criteria Linear Programming Regression (MCLPR) prediction model as the solution. The experiment datasets are provided by a leading Internet company in China, and can be downloaded from track2 of the KDD Cup 2012 datasets. In this paper, Support Vector Regression (SVR) and Logistic Regression (LR) are used as two benchmark models for comparison. The results indicate that MCLPR is a promising model in behavioral targeting tasks.

Keywords: Behavior Targeting; Multiple Criteria Linear Programming Regression (MCLPR); Support Vector Regression (SVR); Logistic Regression (LR);

1. Introduction

Here introduce the paper. With the increasing of Internet users, online advertising becomes an important advertising market and provides a major source of advertising revenues [1]. For web-based businesses, Internet advertising has become a major source of revenue. Internet advertising revenues in the U.S. reached $9.26 billion for the third quarter of 2012, making the quarter the biggest on record, according to the latest IAB Internet Advertising Revenue Report figures released by the Interactive Advertising Bureau (IAB) and PwC US [2]. Major online publishers such as Yahoo!, Microsoft and Google have enthusiastically embraced this

* Corresponding author.Tel.: +86 10 82680697; fax: +86 10 82680697.
E-mail address: fangwangyouxiang@163.com(F. Wang), aapwws@ku.ac.th (W. Suphamitmongkol), wangbo8014@126.com(B. Wang).

doi:10.1016/j.procs.2013.05.103

business model.

The commercial value of advertisement on the Web depends on whether users click on the advertisement. The advertisements click has a significant impact on the Internet industry. It allows Internet companies to identify most relevant ads for each user and improve user experiences. Internet Behavioural targeting (BT) leverages user's online activities to select the ads most relevant to users to display, which is a promising technique to improve the efficiency of online advertising.

There has been a lot of research in Behavioural Targeting. A well-grounded statistical model of BT predicts click-through rate (CTR) of an ad from user behaviour, such as ad clicks and views, page views, search queries etc. The CTR is used in search advertising to rank ads and price clicks. In this paper, we also use the area under the Receiver Operating Characteristics (ROC) curve (AUC) as the evaluation criteria that proposed by track 2, KDD Cup 2012. As we only concern the CTR order of the testing data，the rank of the CTR is used instead of the real value. The predicted AUC score should be higher than 0.5 because 1) the AUC value is between 0.0 and 1.0 and 2) the random guessing value of AUC is 0.5.

Receiver Operating Characteristics (ROC) graph is a useful technique for organizing classifiers and visualizing their performance. ROC graphs are commonly used in medical decision making, and in recent years have been increasingly adopted in the machine learning and data mining research communities. In addition to being a generally useful performance graphing method, they have properties that make them especially useful for domains with skewed class distribution and unequal classification error costs. These characteristics have become increasingly important as research continues into the areas of cost-sensitive learning and learning in the presence of unbalanced classes.

AUC is the Area under the ROC curve, in this paper, which is equivalent to the probability that a random pair of positive samples (clicked ad) and a negative one (unclicked ad) is ranked correctly by using the predicted click-through rate. An equivalent way of maximizing the AUC is to divide each instance into (#click) of positive samples and (#impression-#click)negative samples, and then minimize the pair-wise ranking loss of those samples using the predicted click-through rate[3].

In this paper we utilized Multiple Criteria Linear Programming (MCLP) [3] Regression model to predict the Click-Through rate and to compare it with other two well-known regression methods. The datasets [4] used for testing comes from track2 of the KDD Cup 2012. A major challenge is to create efficient features. Feature creation and selection are the most important steps in solving a supervised learning problem. We compared different methods and then chose two of them to create the features.

The paper is structured as follows. Section 2 reviews related work. Section 3 describes our behaviour data. Section 4 introduces MCLP Regression Data Mining Model and Its Algorithm. Section 5 is the experiment. We conclude the paper in Section 6 with future extended work.

2. Related Work

Much attention has been paid on the advertisement research recently. The best way to maximize the commercial value of advertisements is to display the ads to people who are interested in it. However, there are some issues to be dealt with, such as matching relevant advertisements for a query, ranking of the candidate advertisements, deciding how to display the advertisements on the search result page, click prediction and analysis for the presenting advertisements, and pricing of the advertisements. Several machine learning algorithms such as Logistic Regression , Linear Poisson Regression, Online Bayesian Probit Regression, Support Vector Machines (SVM)[5,6,7,8,9] , and Latent Factor Model have been adopted to predict the clicks of advertisements presented for a query. Since the size of online data is usually huge, online data stream analysis can be very helpful in Behavioral Targeting field [10,11,12].

Behavioral Targeting contains three pricing models, which are Pay-Per-Click (PPC), Pay-Per-Impression (PPI) and Pay-Per-Transaction (PPT). The popular one is PPC. For the PPC model, both the advertiser and the search engine companies wish users to click the advertisements. Therefore, Behavioral Targeting is a good way to solve this problem because it reduces advertiser’s cost and increases search engine companies’ profit

simultaneously. In this paper, we utilized MCLP Regression model to predict the click-through rate (CTR) of ads in a web search engine given its logs in the past and compare it with other two well-known regression methods.

Multiple Criteria Linear Programming(MCLP) is a promising optimization-based classification model [13, 14] and has extended to family toolbox [15]. MCLP has many successful applications including credit card portfolio management [16], credit card risk analysis [17], firm bankruptcy prediction [18, 19], network intrusion detection [20, 21], medical diagnosis and prognosis [22] and classification of HIV-1 mediated neuronal dendritic and synaptic damage [23]. Multi-Criteria Linear Programming Regression (MCLPR) was firstly introduced by Zhang [24], which converted a classification problem to a regression one. The data can be separated into two groups in the way of moving it downward and upward by parameter and then classified by hyperplane to construct regression model. The excellence of MCLPR is its ability to fix the ill-posed condition with limited amount of sample, handling non-linear relationship by kernel function, and giving the global solution if it exists. MCLPR has already proved its performance in many real life datasets.

3. Feature Creation & Selection

In this paper, the training sample comes from track 2 of the KDD Cup 2012 datasets. The training set contains 155,750,158 instances that are derived from log message of search sessions, where a search session refers to an interaction between a user and the search engine. During each session, the user can be impressed with multiple ads, then, the same ads under the same setting (such as position, depth) from multiple sessions are aggregated to make an instance in the datasets. Each instances can be viewed as a vector (#click, #impression, DisplayURL, AdID, AdvertiserID, Depth, Position, QueryID, KeywordID, TitleID, DescriptionID, UserID). It means that under a specific setting, the user (UserID) has been impressed with the ad(AdID) for #impression times, and has clicked #click times of those. In addition to the instances, the datasets also contains token lists of query, keyword, title and description, where a token is a word represented by its hash value. The gender and segmented age information of each user are also provided in the dataset. The test set contains 20,297,594 instances and shares the same format as the training set, except for the lack of #click and #impression. The test set is generated with log messages that come from sessions latter than those of the training set. More detailed information about the datasets can be found in [4]. Feature creation and selection are a major challenge in this paper. We use two different training sets with different feature creation and selection methods in this paper, which are called T-Set-1 and T-Set-2, respectively.

3.1. Feature creation method for T-Set-1

In T-Set-1, the bag of words model was used. This method is frequency-based method that is used to predict the probability of each presented word on a clicked instance based on each feature (tokens). Then, we built the whole feature space by combining the query dictionary and ad dictionary.

3.2. Feature creation method for T-Set-2

Two kinds of features, original features and synthetic features, were used for modeling in this method.

(1) Original Features: The original feature set contains discrete features and continuous features. The discrete features are the unique ID of each ad, advertiser, query, keyword, tile, description, token, gender and age for one user, depth and position of ads, and the displayed URL. The continuous features are the click-through rates of each value of the discrete features. When a discrete feature is being used; the corresponding click-through rate will be activated and adopted as a continuous feature.

(2) Synthetic Feature: First of all, we join any two original discrete features with each other and use them as synthetic features. We also test some 3-tuple features but only the QueryID_AdID_UserID is available. Since most 3-tuple features are too sparse and seldom activated. Secondly, we join the original discrete features with each of the tokens. Position information is added to the original discrete features to generate one 2-tuple position-based feature. Bigram features are also adopted for analyzing the queries, titles and descriptions.

3.3. Normalization

Since the ranges of all the variables' value are significantly different, a linear scaling transformation needs to be performed for each variable. The transformation expresses as below:

$$x_n = \frac{x_i - \min(x_1, K, x_n)}{\max(x_1, K, x_n) - \min(x_1, K, x_n)}$$

where x_n is the normalized value and x_i is the instance value.

4. A Two-Class MCLP Data Mining Model and Its Algorithm

Using MCLP, we can optimize maximizing the minimum distances (MMD) and minimizing the sum of the deviations (MSD) simultaneously, producing better data separation than by linear discriminate analysis. According to the concept of Pareto optimality, we can seek the best trade-off of the two measurements [14, 16]. In this section, we outline the structure of a two-class MCLP model.

Given any two predefined classes {G: Good and B: Bad} for a datasets, given training samples $T_n = \{A_1, A_2, ..., A_n\}$, where n is the total number of records in the training sample. Each training instance A_i has r attributes. This data mining model is used to determine the coefficients for an appropriate subset of the variables, denoted by $X = (x_{1,...,} x_r)$, and a boundary value b to separate two classes: G (Good) and B (Bad) with minimizing the overlapping; that is, if $A_i X < b, A_i \in G$ and if $A_i X > b, A_i \in B$, where A_i is the vector value of the subset of variables from the database and the symbol "$\in$" means "belongs to". Note that when $A_i X = b$, A_i belongs to either G or B. The geometric meaning of the model is shown in Fig.1. (a).

To measure the separation of G and B, we define:

α_i = the overlapping of a two-class boundary for case A_i (external measurement);

β_i = the distance of case A_i from its adjusted boundary (internal measurement);

We use (☆) to represent A_i of Good and (•) to represent A_i of Bad. Fig.1.(a) shows that our goal is to minimize the sum of α_i and maximize the sum of β_i simultaneously. As a result, two groups of data represented in Fig.1.(a) will be pulled away. Therefore, this model can be written as:

$$Minimize \sum_i \alpha_i \; and \; Maximize \sum_i \beta_i$$

Subject to:

$$A_i X = b + \alpha_i - \beta_i, A_i \in G,$$
$$A_i X = b - \alpha_i + \beta_i, A_i \in B,$$

where A_i are given, X and b are unrestricted, and α_i and $\beta_i \geq 0$.

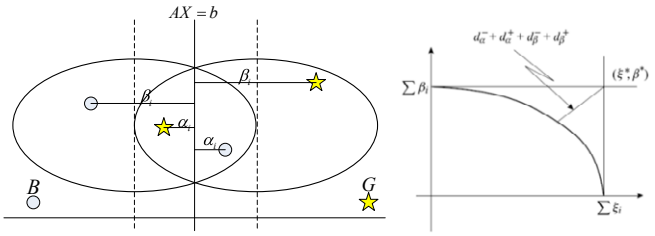

Fig.1. (a) geometric meaning of MCLP; (b) geometric meaning of compromise solution of MCLP

To facilitate the computation, the compromise solution approach [25] can be employed to reform the above model so that we can systematically identify the best trade-off between $-\sum_i \alpha_i$ and $\sum_i \beta_i$ for an optimal solution. The two-class MCLP model has evolved to the following model:

$$\begin{aligned}
&Minimize\ d_\alpha^- + d_\alpha^+ + d_\beta^- + d_\beta^+ \\
&Subject\ to: \\
&\alpha^* + \textstyle\sum_i \alpha_i = d_\alpha^- - d_\alpha^+, \\
&\beta^* - \textstyle\sum_i \beta_i = d_\beta^- - d_\beta^+, \\
&A_i X = b + \alpha_i - \beta_i,\ A_i \in G, \\
&A_i X = b - \alpha_i + \beta_i,\ A_i \in B, \\
&\alpha_i, \beta_i, d_\alpha^-, d_\alpha^+, d_\beta^-, d_\beta^+ \geq 0
\end{aligned}$$

where $A_i, \alpha^*, and\ \beta^*$ are given, $X\ and\ b$ are unrestricted, and $\alpha_i, \beta_i, d_\alpha^-, d_\alpha^+, d_\beta^-, d_\beta^+ \geq 0$. The geometric meaning of compromise solution of MCLP is shown in Fig.1.(b).

5. Multi-Criteria Linear Programming Regression (MCLPR) Data Mining Model and Its Algorithm

MCLPR is an extension of MCLP, which proposed by Shi [3] and now is widely used in data mining field.

Consider a data set of a regression problem:

$$T = \left\{(x_1^T, z_1), (x_2^T, z_2), ..., (x_n^T, z_n)\right\},$$

where $x_i \in R^r$ are the input variables, and $z_i \in R$ is the output variable, which can be any real number. Define the G and B as "Good" (-1) and "Bad" $(+1)$ (a binary case), respectively. Then the corresponding S_{MCLP}^- and

S^{+}_{MCLP} datasets for MCLP regression model are constructed. With these datasets, the MCLP regression model is formalized as follows:

$$
\begin{aligned}
&Min\sum_{i=1}^{n}(\xi_i+\xi_i^{'})-\sum_{i=1}^{n}(\beta_i+\beta_i^{'})\\
&Subject\ to\text{:}\\
&\chi_{11}w_1+\ldots+\chi_{1r}w_r+(y_1+\xi)w_{r+1}=b-\xi_1+\beta_1 \quad for\ all\in G\\
&\left.\begin{array}{l}\chi_{n1}w_1+\ldots+\chi_{nr}w_r+(y_n+\xi)w_{r+1}=b-\xi_n+\beta_n\\ \chi_{11}w_1+\ldots+\chi_{1r}w_r+(y_1-\xi)w_{r+1}=b+\xi_i^{'}-\beta_i^{'}\end{array}\right\}for\ all\in B\\
&\chi_{n1}w_1+\ldots+\chi_{nr}w_r+(y_n-\xi)w_{r+1}=b+\xi_n^{'}-\beta_n^{'}\\
&\qquad\qquad \xi,\xi^{'},\beta,\beta^{'}\geq 0
\end{aligned}
$$

where $(x_i,\ y_i), i=1,\ldots n$, denote the training data, where x_i is an input and $y_i \in R$. Then, the training data are transformed into two groups, Good (G) and Bad (B) by predetermined parameter (precision error). The new training data are constructed as follows:

$$
\begin{aligned}
&Good\ Samples: D^{+}=\{((x_i,y_i+\xi),+1), i=1,\ldots,n\}\\
&Bad\ \ Samples: D^{-}=\{((x_i,y_i-\xi),-1), i=1,\ldots,n\}
\end{aligned}
$$

For facilitate the computation and improve regression result, we also apply the compromise approach mentioned in MCLP classification [25]. In this paper, $\xi^{*}=-10^{-6}$ *and* $\beta^{*}=10^{6}$ are used as the ideal value of each parameter.

6. Experiment

In this paper, we used two files (Dataset1, Dataset2) for training, which were generated and selected by the methods in section 3, respectively. Dataset1 was the subset of the T-Set-1, Dataset2 was the subset of the T-Set-2 and Principal Component Analysis (PCA) was then used for feature selection. Dataset1 contained 639999 records for training and 68447 records for testing. Dataset2 contained 20000 records for training and 10000 records for testing. The baseline models were Support Vector Regression (SVR) and Logistic Regression (LR). In this paper, we also use the area under the Receiver Operating Characteristics (ROC) curve (AUC) as the evaluation criteria that proposed by track 2, KDD Cup 2012 [4].

6.1. Result

As shown in table 1 and 2, The values of AUC of the three models were very close to each other and the training results on Dataset2 were better than those on Dataset1. The best result of AUC, which was trained by Support Vector Regression (SVR) on Dataset2, was 0.7913. Multiple Criteria Linear Programming Regression (MCLPR) was the second one in both of the two datasets, and Logistic Regression (LR) was the last one.

Comparing with the results of the top four winner of track 2, KDD CUP 2012 [26], all the results were in reasonable agreement in this paper. The results of the experiment demonstrate that MCLPR is also a good alternative method in the research field of Behavioral Targeting. Comparing with the traditional mathematical tools in classification, such as neural networks, decision tree, and statistics, MCLP is simple and direct, free of the statistical assumptions, and flexible by allowing decision makers to play an active part in the analysis.

Table 1. Comparisons of dataset 1 with different models.

Model	AUC
SVR	0.7888
MCLPR	0.7829
LR	0.7821

Table 2. Comparisons of dataset 2 with different models.

Model	AUC
SVR	0.7913
MCLPR	0.7878
LR	0.7826

Figure 3 and 4 showed the ROC curves of SVR (red line), MCLPR (green line) and LR (yellow line). The horizontal axis represents the False Positive rate and the vertical axis represents the True Positive rate. SVR obtained the best performance among all of the three models and in both of two datasets. Multiple Criteria Linear Programming Regression (MCLPR) was the second one in both of the two datasets, and Logistic Regression (LR) was the last one in both of the two datasets.

For the curves, the closer to the upper left corner is the better one, which means the True Positive rate is higher. Usually, the diagonal was used as the baseline. The ROC curves should be above the diagonal. The ROC curves in the figures demonstrate that our experiment results are meaningful in Behavioral Targeting field. The biggest contribution of this paper is that a new model (Multiple Criteria Linear Programming Regression) was proposed and proved to be useful and valuable in Behavioral Targeting field.

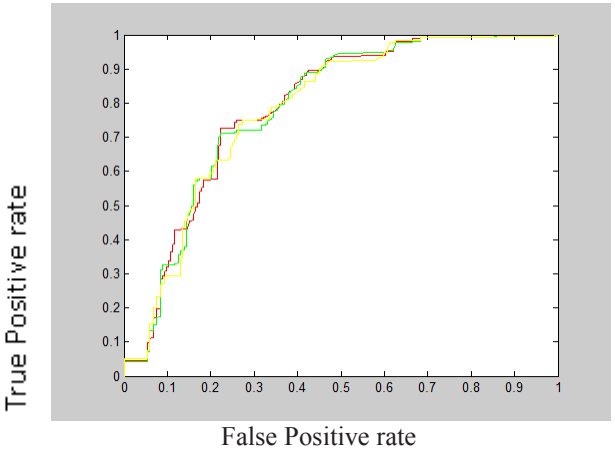

Fig. 3. ROC curves of SVR (red line), MCLPR (green line) and LR (yellow line) on DataSet1.

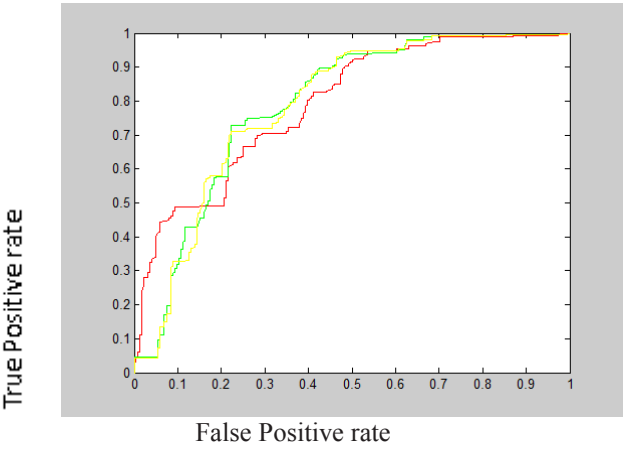

Fig. 4. ROC curves of SVR (red line), MCLPR (green line) and LR (yellow line) on DataSet2.

7. Conclusions

In this paper, a multiple criteria linear programming Regression (MCLPR) algorithm has been proposed to predict the advertisement Click-Through Rate, which is new in Behavioral Targeting field. The experiment results demonstrate that MCLPR is an efficient method in predicting Click-Through Rate. In the future, we will extend the method by integrating with other models (ensemble model) to improve the prediction result. As many potential customers exist on the Internet, the User's social data will be added into the training sample to solve the data sparse problem.

Acknowledgement

This work has been partially supported by grants from National Natural Science Foundation of China (Nos.70921061, 11271361, 71201143, 61003167), the CAS/SAFEA International Partnership Program for Creative Research Teams, Major International (Ragional) Joint Research Project (No.71110107026), the President Fund of GUCAS.

References

[1] J. Li, P. Zhang, Y. Cao, P. Liu and L.Guo, Efficient Behavior Targeting Using SVM Ensemble Indexing. In Proceedings of the 12th IEEE International Conference on Data Mining (ICDM-12), December 10-13, 2012, Brussels, Belgium.

[2] http://www.iab.net/about_the_iab/recent_press_releases/press_release_archive/press_release/pr-121912.

[3] Y. Shi (2001) Multiple Criteria and Multiple Constraint Level Linear Programming: Concepts, Techniques and Applications, World Scientific Publishing Co.

[4] http://www.kddcup2012.org/c/kddcup2012-track2.

[5] Y. Tian,Y. Shi, X. Liu, Recent advances on support vector machines research, Technological and Economic Development of Economy 18(1) 5--33.

[6] Z. Qi, Y. Tian, Y. Shi, Robust twin support vector machine for pattern classification, Pattern Recognition, 2013, 46(1): 305-316.

[7] Z. Qi, Y. Tian, Y. Shi, Laplacian twin support vector machine for semi-supervised classification, Neural Networks, 2012, 35:46-53.
[8] Z. Qi, Y. Tian, Y. Shi, Twin support vector machine with Universum data, Neural Networks, 2012, 36C:112-119.
[9]Z. Qi, Y. Tian, and Y. Shi, Structural Twin Support Vector Machine for Classification, Knowledge-Based Systems, 2013, DOI: 10.1016/j.knosys.2013.01.008.
[10] P. Zhang, X. Zhu, Y. Shi, L. Guo, and X. Wu, "Robust Ensemble Learning for Mining Noisy Data Streams". Decision Support Systems, Vol. 50(2), 2011, pages: 469-479.
[11] P. Zhang, J. Li, P. Wang, B. Gao, X. Zhu, and L. Guo, "Enabling Fast Prediction for Ensemble Models on Data Streams". In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-11), August 21-24, 2011, San Diego, CA, USA.
[12] P. Zhang, B. Gao, P. Liu, Y. Shi, and L. Guo, "A Framework for Application-Driven Classification of Data Streams". Neurocomputing 92 (2012), 170-182.
[13] Y. Shi, et al., Data mining in credit card portfolio management: a multiple criteria decision making approach. Lecture notes in economics and mathematical systems, 2001: p. 427-436.
[14] Y. Shi, Y. Peng, and W. Xu, Data mining via multiple criteria linear programming: Applications in credit card portfolio management. International Journal of Information Technology and Decision Making, 2002. 1(1): p. 131-151.
[15] A. Li, et al., A Fuzzy Linear Programming-Based Classification Method. International Journal of Information Technology & Decision Making, 2011. 10(06): p. 1161-1174.
[16] Y. Shi, et al., Data mining in credit card portfolio management: a multiple criteria decision making approach. Lecture notes in economics and mathematical systems, 2001: p. 427-436.
[17] Y. Peng, et al., A Multi-criteria Convex Quadratic Programming model for credit data analysis. Decision Support Systems, 2008. 44(4): p. 1016-1030.
[18] Kwak, W., Y. Shi, and J.J. Cheh, Firm bankruptcy prediction using multiple criteria linear programming data mining approach. Advances in Investment Analysis and Portfolio Management, 2006(2): p. 27-49.
[19] Kwak, W., et al., Bankruptcy prediction for Japanese firms: using Multiple Criteria Linear Programming data mining approach. International Journal of Business Intelligence and Data Mining, 2006. 1(4): p. 401-416.
[20] G. Kou, et al., Multiple criteria mathematical programming for multi-class classification and application in network intrusion detection. Information Sciences, 2009. 179(4): p. 371-381.
[21] Y. Shi, et al., A Multiple-Criteria Quadratic Programming Approach to Network Intrusion Detection, in Data Mining and Knowledge Management. 2005, Springer Berlin / Heidelberg. p. 145-153.
[22] Z. Zhang, Y. Shi, and G. Gao, A rough set-based multiple criteria linear programming approach for the medical diagnosis and prognosis. Expert Systems with Applications, 2009. 36(5): p. 8932-8937.
[23] J. Zheng, et al., Classification of HIV-I mediated neuronal dendritic and synaptic damage using multiple criteria linear programming. Neuroinformatics, 2004. 2(3): p. 303-326.
[24] D. Zhang, Y. Shi, Y. Tian,M. Zhu, A class of classification and regression methods by multiobjective programming. Frontiers of Computer Science in China , 2009. 2(3): p. 192-204.
[25] H. Lee, Y. Shi, J. Stolen, Allocating data files over a wide area network: goal setting and compromise design, Information & Management, 1994. 2(26) : P. 85–93.
[26] http://www.kddcup2012.org/workshop.